In collaboration with In Concert

Nurmikko-fuller, Terhi; Dix, Alan; Weigl, David M.; Page, Kevin R.

DOI:
10.1145/2970044.2970049

Document Version
Peer reviewed version

Citation for published version (Harvard):
Nurmikko-fuller, T, Dix, A, Weigl, DM & Page, KR 2016, In collaboration with In Concert: Reflecting a digital library as linked data for performance ephemera. in *DLfM 2016: Proceedings of the 3rd International Workshop on Digital Libraries for Musicology.* Association for Computing Machinery , pp. 17-24. https://doi.org/10.1145/2970044.2970049

In Collaboration with *In Concert*: Reflecting a Digital Library as Linked Data for Performance Ephemera

Terhi Nurmikko-Fuller
Oxford e-Research Centre
University of Oxford, UK
terhi.nurmikko-fuller@oerc.ox.ac.uk

Alan Dix
University of Birmingham, UK
and Talis, Birmingham, UK
alanjohndix@gmail.com

David M. Weigl
Kevin R. Page
Oxford e-Research Centre
University of Oxford, UK
{david.weigl, kevin.page}@oerc.ox.ac.uk

ABSTRACT

Diverse datasets in the area of Digital Musicology expose complementary information describing works, composers, performers, and wider historical and cultural contexts. Interlinking across such datasets enables new digital methods of scholarly investigation. Such bridging presents challenges when working with legacy tabular or relational datasets that do not natively facilitate linking and referencing to and from external sources. Here, we present pragmatic approaches in turning such legacy datasets into linked data.

InConcert is a research collaboration exemplifying these approaches. In this paper, we describe and build on this resource, which is comprised of distinct digital libraries focusing on performance data and on concert ephemera. These datasets were merged with each other and opened up for enrichment from other sources on the Web via conversion to RDF. We outline the main features of the constituent datasets, describe conversion workflows, and perform a comparative analysis. Our findings provide practical recommendations for future efforts focused on exposing legacy datasets as linked data.

CCS Concepts

•Information systems → Resource Description Framework (RDF); •Applied computing → Digital libraries and archives; *Performing arts;* **•Theory of computation →** *Data provenance;* **•Computing methodologies →** *Ontology engineering;*

Keywords

linked data, RDF, concert ephemera, performance metadata, workflows, batch and live processing

1. INTRODUCTION

In this paper we describe the production of a linked data digital library (as defined by Bainbridge, et al.[3]) focused on historical musical performances, and on concert ephemera, combining several constituent data sources, each originally structured as a distinct tabular or relational dataset. We report on accomplished prior aims [11], and describe the methodological workflows used to expose these legacy datasets as linked data[1] , serving as practical guidance for projects with similar goals, and allowing the process to be replicated to verify our results.

The paper is divided into two threads. Each focuses on a separate constituent dataset, discussing the implications of the data, and detailing the conversion workflows. We note the effect of human factors, and relate discovered benefits and challenges of both live and batch processing. These threads are then brought together for a comparative analysis, resulting in a list of recommendations, findings, and views to future work.

DLfM '16, August 12 2016, New York, USA

ISBN 978-1-4503-4751-8/16/08...$15.00

DOI: http://dx.doi.org/10.1145/2970044.2970049

2. BACKGROUND

In Concert: Towards a Collaborative Digital Archive of Musical Ephemera (henceforth "*InConcert*")[2], is a collaborative sub-project of the UK Arts and Humanities Research Council funded *Transforming Musicology*[3]. It examines bibliographical and performance metadata sourced from concert ephemera such as programmes, bills, reviews, and advertisements from historical newspapers and periodicals.

InConcert bridges the data from *Calendar of London Concerts 1750 - 1800* (LC18)[18], and *19th-century London Concert Life (1815 - 1895)* (LC19)[4], both described in greater detail below. The inclusion of Optical Character Recognition (OCR) data from an additional resource, the *Concert Programme Exchange (1901 - 1914)* (CPE)[1][4] was considered, but the nature of the OCR rendered it unsuitable within the available time constraints of the project. Instead,

[1]The linked data approach employs the Resource Description Framework (RDF), a standard model for online data exchange that specifies data instances and relations using Universal Resource Identifiers (URIs). A set of two instances joined by a relation is referred to as a *triple*. This approach allows the meanings of the relationships between the data to be interpretable by both humans and machines, and enables the linking of this information to external datasets, embedding it within a wider web of knowledge, making it discoverable and promoting reuse in other contexts [6].

[2]http://inconcert.datatodata.com

[3]http://transforming-musicology.org/

[4]CPE focuses on early 20th century data, capturing the exchange of seasonal concert programmes between major European concert venues.

OCR data of the British Musical Bibliography (BMB)[7] was transformed via a mixture of automated scripts correcting original OCR, and exception files.

The aim of *InConcert* was to build a digital library containing various types of performance datasets that would enable prosopographical analyses, and aid in the understanding and re-imagining of the digital archiving process. The data platform is based on a core information system design principle, which regards the original data as the 'Golden Copy' – subsequent processing may cache data, augment it, or represent it in different formats, but the integrity of the original data is paramount. It deviates from more common information architectures, which regard the central repository as definitive.

This has led to a number of design decisions:

- *continuous update* – workflows from the original to the processed data should be reproducible,
- *specify rather than transform* – data is left as close to is original form as possible, with additional specifications to reveal the semantics of the data.

Available information ranges from simple, raw data derived from OCR processes to richly interpreted and highly structured data with multiple layers of linkage and verification. It has been produced using methodologies combining automated processes with the input of expert musicologists.

3. RELATED WORK

The use of semantic web technologies to support study of digital music objects has previously been previously implemented [9] [10] and successfully applied to other projects under the auspices of *Transforming Musicology* [8]. Projects such as *SALAMI: Structural Analysis of Large Amounts of Music Information* [5] and *RISM:Répertoire International des Sources Musicales*[5] are illustrative of recent projects with similar research agendas, whilst the *Répertoire International de Littérature Musicale*[6] exemplifies on-going work in the field of ontology design for musicological data. Other complementary ontologies are currently under development within the larger context of *Transforming Musicology* on the nature of leitmotifs [15], as well as an extension or revision of the CHARM [26] ontology by [16].

We made us of a number of existing ontologies, namely the interconnected Music [23], Event [22], and Timeline [21] Ontologies, as well as Schema[7], and the bibliographic metadata ontologies of Bibframe[8], and FaBiO [24]. These ontologies were insufficient to completely map all available *InConcert* data, and whilst they contributed extensively to the used underlying structure, some new ontological development formed part of the workflow.

In previous work by one of the authors, similar techniques were used to build a working set of linked data [13]; indeed this can be seen as a distributed extension of traditional heuristic-led algorithms or early 'snowball sample' web search ranking [17].

[5]https://opac.rism.info/metaopac/start.do?View=rism
[6]http://www.rilm.org/
[7]http://schema.org/
[8]https://www.loc.gov/bibframe/

4. CALENDAR OF LONDON CONCERTS 1750-1800 (LC18)

LC18 is openly available in tabular format (.csv, .xls) under Creative Commons Attribution Non-commercial Share Alike. Both documentation and the data are accessible [18].

4.1 Source Data

A stable dump of a pre-existing relational database, the main files in LC18 are tabular .csv transformed to JSON and imported into a noSQL database.[9] The categories of LC18 data are shown in the sample record listed in Table 1.

Temporal information is available in three categories, including: "Date" (row 3), capturing the numeric value for the historical date; "Day", which has additional information including an abbreviation of the day of the week; and "Time" capturing the performance time on a 24-hr clock.

Much of the information is expressed via acronyms. Geographical data is captured via "Place", and represented as a short (often a two or three letter) initialisation, such as "CG", referring to a specific venue within London (in this case, the Covent Garden Theatre). The performance title may also contain the venue acronym and the performance type. The latter is additionally captured under "Type".

"Price" contains complex information about the seating types and the respective prices of different categories of seat, but is expressed via a single, undifferentiated text string. For the performance described in Table 1, four different types of tickets were sold: the pit, and the boxes, as well as those for the first and second floor galleries. For the latter two, the prices are different and the seat-price pairing is explicitly stated - for the pit and the box, the seats are the same price ("10s 6d", i.e. "Half a guinea" [18]).

The remainder of the table captures bibliographic information regarding the sources in which the performance was recorded (Programme, row 9), advertised (row 10), and in some cases, reviewed (row 11).

Accessing information in this tabular format, with cross-referencing to the available documentation, is a feasible option for human experts, but is difficult for software agents and digital tooling. Many categories (such as price and performance type) have embedded implicit information, which needs to be explicitly addressed and captured in the final RDF if these categories are to be addressed by new research questions that are made possible by the conversion of this data using semantic technologies. These relationships and new entities form part of the ontological structure designed to capture LC18 data (Fig 1).

4.2 Ontological Modeling

Capturing LC18 data as RDF necessitated the development of an underlying ontological structure. This 'bottom-up' model was extensively based on the tabular structure, where for the large part classes are directly aligned with information types of the original data: classes in the ontology (e.g. InC:Title, as illustrated by Fig 1) contain as their instance data the content of the column of "Title" from the original flat table (represented as row 5 in Table 1).

During the iterative design process the requirements of the data became more apparent, and possibilities for further

[9]noSQLite is a lightweight noSQL library designed to run on standard LAMP-stack without additional daemons or sysadmin installation.

Table 1: Sample record: Calendar of London Concerts 1750-1800 (*LC18*)

	Data	Content	Expanded acronym
1	No	15	
2	Date	1750_03_14	
3	Day	Wed 14 Mar 1750	
4	Place	CG	Covent Garden Theatre
5	Title	CG ORATORIO 50 [4TH]	
6	Type	OS	Oratorio Series
7	Time	1830	
8	Price	PB 10s 6d; FG 5s; SG 3s 6d	Pit and Boxes, First (lower) gallery, Second (upper) gallery
9	Programme	handel o ˆJUDAS_MACCABAEUS	
10	Advert	GA	General Advertiser; (Parker's) General Advertiser
11	Review	See Deutsch–Handel 683-4	
12	Notes	LS; Burrows–Harris 267.]	*The London Stage 1660–1800: Part 4*, ed. G.W. Stone (Carbondale, 1962); Part 5, ed. C.B. Hogan (Carbondale, 1968) ; Donald Burrows and Rosemary Dunhill, *Music and Theatre in Handel's World: The Family Papers of James Harris 1732–1780* (Oxford, 2002)

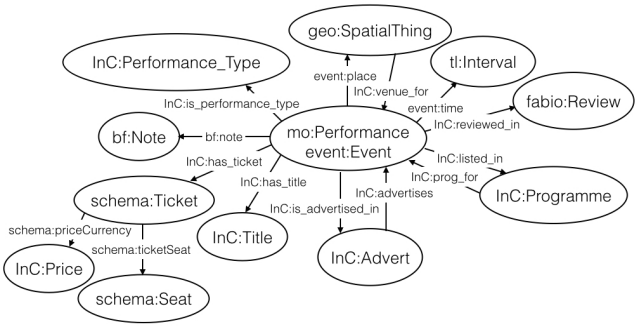

Figure 1: Ontology for LC18

semantic enrichment of the data were identified. Certain controlled descriptors that have the potential to be semantically meaningful are currently captured as unstructured plain-text (expressed using the generic rdfs:label predicate); for instance, the content of row 6 in the sample record (Table 1) is a plain-text rdfs:label for lnC:Performance_Type, and the information embedded within that string is only available in a human-readable format.

Future iterations of the ontological structure will further expose such controlled terms as distinct entities. In this case, it would be achieved by explicitly representing (or indeed *reifying*) Oratorio Series as a subclass of the Music Ontology's mo:Performance class. Similar finer granularity of semantically meaningful data instances could be explicitly declared for other aspects of the data, such as seat type (e.g. Pits cf. First Gallery; see row 8).

The adequate and complete mapping of this type of information necessitates the development of a fully-fledged ontology of performance data, which has been beyond the scope of the current research project.

4.3 Implementation: Web-Karma

The conversion of the LC18 data[10] from the tabular (.csv) to the Turtle[11] (.ttl) format was carried out using Web-Karma[12], an Open Source software tool developed at the University of Southern California. Our workflow (Fig 2), which consisted of mapping between our imported (OWL[13]) ontological structure (Fig 1) and the data (Table 1), and was completed using the tool's graphical user-interface, is a reiteration of the processes employed in an earlier project focusing on the alignment and linking of bibliographic metadata from two large digital library corpora [19].

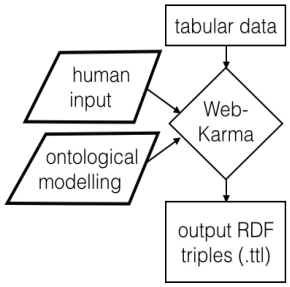

Figure 2: LC18 conversion workflow

The limitations of Web-Karma necessitated minimal data editing, namely the replacement of colons and semi-colons in strings. The process requires the prior design and implementation of an ontological structure, and a high degree of familiarity with the data. Advantages include the production of high-quality triples with little subsequent correction, adhering to best practice by minimizing the use of blank nodes [14]. This workflow produced over 95,000 triples.

[10] http://datatodata.com/in-concert/LC18/list.php?type=concerts
[11] http://www.w3.org/TR/turtle/
[12] http://usc-isi-i2.github.io/karma/
[13] http://www.w3.org/OWL/

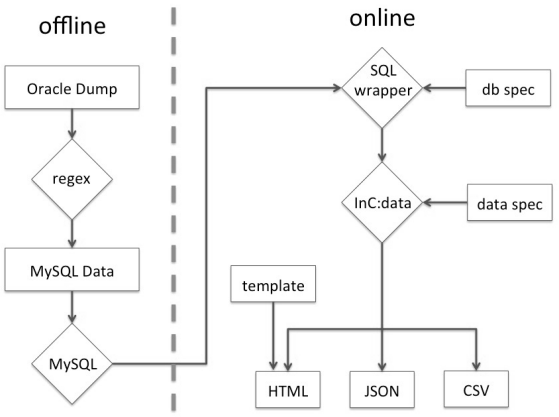

Figure 3: Workflow to create the dataset for LC19

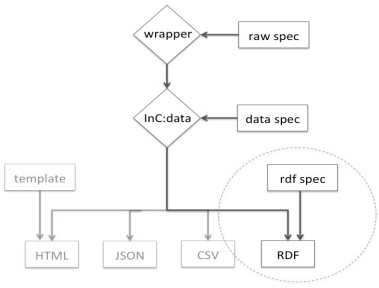

Figure 4: New InConcert data infrastructure including online generation of RDF

5. 19TH CENTURY LONDON CONCERT LIFE 1815 - 1895 (LC19)

Unlike LC18, which focuses on performance metadata, LC19 is currently a corpus of concert ephemera. The data is largely based on capturing bibliographical metadata of pamphlets, newspapers, and other printed material that captures historical information about performances, people, and locations. The intention is that a separate interpretation activity will use this data to construct an authoritative catalogue of concerts more closely paralleling those in LC18. One of the aims of *InConcert* is to make this interpretation task easier.

The data infrastructure supports the outputs of any single item, list, or search in a number of different formats: HTML, .csv, and JSON (Fig 3). As an outcome of the work presented here, RDF is added to this list of outputs, based on two distinct workflows: an ontological mapping using D2R[14], and a JSON-LD exemplar (both described below; see also Fig 4). Where only incomplete (or no) mapping specifications were available, default transformations were applied. The resultant RDF consists of fully linked data bound to persistent URIs. It is produced on-the-fly from the original data, adhering to our *specify rather than transform* design decision.

5.1 Source Data

LC19 project data is contained in a MySQL database, itself based on a legacy Oracle dump. The relational structure of the dataset is provided as a cluster of some twenty files, consisting largely of the core tables, but including earlier, now superfluous versions. Instance-level data is not publicly accessible, but was shared confidentially within the project. Due to differences in Oracle and MySQL schema syntax, the original dump had to be manually transformed using regular expressions and special cases imported into MySQL. Large data (images and text files) are not included in the database, but are available separately in a web-site dump linked via partial paths in specific database columns.

[14]http://d2rq.org/d2r-server

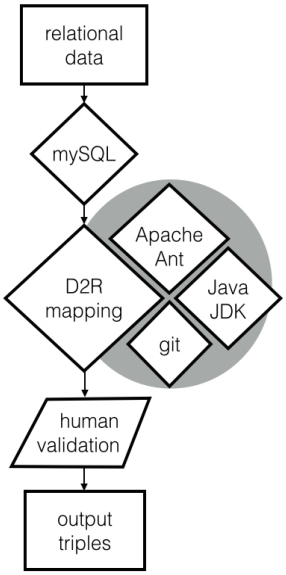

Figure 5: D2R workflow for LC19

```
map:CONCERTS a d2rq:ClassMap ;
        d2rq:dataStorage map:database ;
        d2rq:uriPattern "CONCERTS/@@CONCERTS.CONCERT_ID@@" ;
        d2rq:class vocab:CONCERTS ;
        d2rq:classDefinitionLabel "CONCERTS" .

map:CONCERTS_label a d2rq:PropertyBridge ;
        d2rq:belongsToClassMap map:CONCERTS ;
        d2rq:property rdfs:label ;
        d2rq:pattern "CONCERTS #@@CONCERTS.CONCERT_ID@@" .
```

Figure 6: Sample D2R mappings (.ttl format) derived from the LC19 MySQL database

5.2 Implementation: D2R

LC19 data was encoded as RDF through two workflows. The first took advantage of existing experience within the wider *Transforming Musicology* project, and used D2R to automatically generate semantic mappings producing RDF triples from the relational data in the MySQL database [8]. D2R installation has dependencies on git[15], Apache Ant[16], and Java JDK[17], but requires little user time beyond the initial installation (Fig 5).

The RDF triples are generated automatically in a two-

[15]http://git-scm.com/
[16]http://ant.apache.org/
[17]http://www.oracle.com/technetwork/java/javase/downloads/index.html

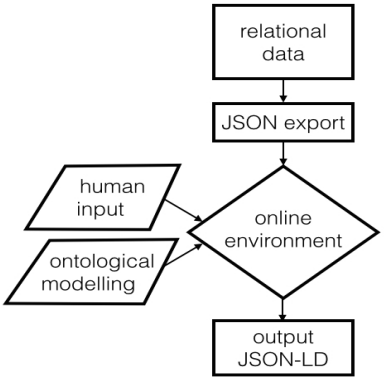

Figure 7: JSON-LD workflow for LC19

```
"surname": "foaf:familyName",
"first_names": "foaf:givenName",
"title": "foaf:title",
"gender": {
  "property_uri": "foaf:gender",
  "value_map": { M": "male", "F": "female" }
},
"country_birth": {
  "trim": "both",
  "parse": [{
   "regex": "\/([A-Za-z]+)\\s*\\(([A-Za-z]+)\\)\/",
   "parts": {  country": 1, area": 2  }
    },
    { "default": "country" }
  ]
},
"country_birth.country": {
  "property_uri": "schema:BirthPlace",
  "value_pattern":
   "http:\/\/dbpedia.org\/page\/{country_birth.country}",
  "rdftype": "uri"
},
"birth_year": {
  "property_uri": "schema:birthDate",
  "datatype": "xsd:date"
},
```

Figure 8: Fragment of JSON-LD context specification, mapping simple fields to property URIs. Unmapped fields are given a default namespace and mapping.

fold process; the first iteration captures the data mappings (see sample in Fig 6), and the second produces instance-level triples. These are all mapped to the vocab: namespace, but can be changed using a DELETE-INSERT SPARQL query to match other suitable ontologies such as as FOAF[18]: consider for example foaf:Person or foaf:Agent as a replacement for vocab:PEOPLE. This particular conversion is supported by the complementary overlap between LC19 data relationships and FOAF predicates, e.g. foaf:familyName to capture vocab:PEOPLE_A_SURNAME.

This approach resulted in a repetition of classes and their labels. For example, the class for vocab:CONCERTS, also has the label (a plain-text string) "CONCERTS" (Fig 6).

[18]http://xmlns.com/foaf/spec/

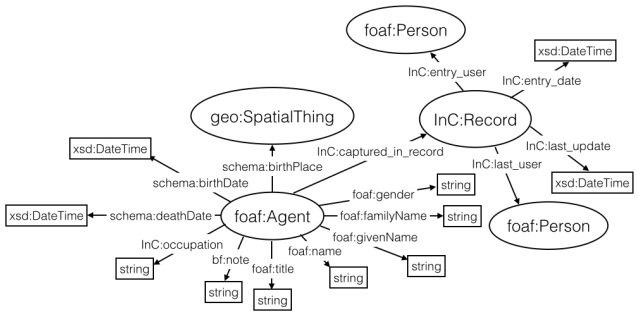

Figure 9: Person section of the LC19 ontology

5.3 Implementation: JSON-LD

In addition to the D2R workflow described above, a simple case-study example focusing on person data was carried out using JSON-LD and an online development environment[19] (Fig 7). LC19 had data exports available in different formats, including JSON. This stage of the project was an investigation to assess how easily and successfully JSON-LD could be produced in the conversion process. The workflow is based on a singular example and is not currently scalable, although JSON-to-RDF conversion tools that use JSON-LD for larger-scale processing have previously been described in a digital musicology context [25][20].

The instance level data was mapped to a number of ontologies (Fig 8) capturing person biography (foaf:familyName, foaf:givenName, foaf:title, foaf:gender and schema:BirthPlace, schema:birthDate), bibliographical information (making use of the Bibframe[21] ontology's bf:note and bf:Note, as well as FaBio[20] classes and predicates) and additional metadata (purpose-built *In Concert* ontology (e.g. InC:last_user)). This approach requires greater effort in terms of user-input than the D2R workflow (section 5.2), but the generated RDF triples require a lesser degree of post-hoc editing.

A mapping was created to ensure the ontological validity of the data. This highlighted an interesting aspect, which was not immediately clear from the D2R generated RDF triples: tables of the relational database could hold information about semantically different entities. In the case of PERSON, the instance-level MySQL data describes information about the person along with metadata detailing the person's *record* in the database. Thus, the data might describe a person with a birth date in the early 1800s, but also capture the entry_date of the database record from 1998. To adequately capture this detail in the RDF output, an entity corresponding to the notion of the Record must be created, and it is to this entity that details relating to the record's provenance, e.g. user identifiers, creation and edition dates are attached (Fig 9).

6. EXPERT INPUT FOR LC18 AND LC19

The need for human activity and expertise has been highlighted throughout the largely data-oriented narrative we have presented so far. These individual points of input can be seen as part of a richly interconnected mutual scaffolding of data development and human interactions.

[19]http://json-ld.org/playground/
[20]https://github.com/musicog/json-ld-scraper
[21]http://bibframe.org/

The *musicologists* collected the original data, which itself required complex 'traditional' data modelling for the SQL database. An early task of the *HCI/IT expert* was to produce versions of the human-readable formats of tabular and SQL data – these were essential to establish a shared point of understanding between the human agents involved, providing the *HCI/IT expert* sufficient insight to enter into richer discussions with the *musicologists* about their data.

The *HCI/IT expert* could then create revised views of the data, facilitating further discussions where the *musicologists* saw patterns in their own data of which they had been previously unaware, and enabling the creation of specialized interfaces for the *musicologists* to enrich the datasets.

The *ontologist* used human-readable data views to gain sufficient understanding to enter into more detailed discussions with the *HCI/IT expert*, guided by previous interactions with the *musicologists*, informing the ontology modeling, batch RDF creation, and JSON-LD exemplar (Fig 8). This enabled the *HCI/IT expert* to add live RDF .ttl as a data option to the *InConcert* platform.

Each stage of human interpretation influenced the triples that were generated. The *ontologist*'s understanding of the data structures for the MySQL directly fed into the reclassification RDF triples generated by D2R. The *ontologist*'s understanding of the JSON data is directly captured in the production of the JSON-LD, and in return feeds back to the *database administrator*, the descriptor files, and ultimately contributes to the final version of the produced linked data.

7. COMPARATIVE ANALYSIS

Although both are part of *InConcert*, LC18 and LC19 differ from each other as datasets. There are three main points of difference, two of which have directly influenced the decisions for workflow use. Firstly, the datasets contain related, non-identical data: one focuses on performance, the other on bibliographical material. The two do not overlap on the temporal spectrum, although the close proximity of the end of LC18 to the start of LC19 means there are instance-level parallels contained in the data. The second difference is in the data structure, with LC18 accessible as tabular data, and LC19 via JSON export of a relational database dump. The third difference between the two is one of access: LC18 data is openly available, whereas LC19 is not.

7.1 Comparing Data

The differences in the data content (i.e. performance metadata vs. bibliographical metadata) has not influenced the choice in workflow or RDF output. To ensure that the two can be linked at class level, we have deliberately used the same ontologies to map each dataset – e.g. foaf:Agent, foaf:Person, and geo:SpatialThing. This was done to ensure schema-level linking between the ontological structures, and also supports linking via shared instance-level entities.

7.2 Comparing Implementations

The different approaches to data structure had the greatest effect on our workflow choices. The chosen tools are tied to the corresponding data structures: Web-Karma cannot be used with relational data, nor can D2R be used to convert the mappings of a tabular dataset into RDF. The JSON-LD workflow (see section 5.3) could be used for both datasets, but requires additional programming tasks to be scalable.

These approaches to RDF conversion are suited to stable datasets. The original LC18 data was produced using exhaustive methods, with occasional updates as new source data came to light – similarly, the RDF conversion was carried out on data that is not expected to undergo extensive updating or change in the foreseeable future. The LC19 data is more likely to be dynamic, as its role within the wider *InConcert* context is to enable scholarly editing.

Changes to either dataset will necessitate a repetition of the RDF production workflows. This is particularly significant for two reasons: firstly, the project may require periodic effort, and secondly, this reiteration of workflows runs counter to the "Golden Copy" design principles (outlined above), as here is potential that the produced RDF will begin to be viewed as the canonical or definitive version. This may be further exacerbated in situations where the RDF is copied into different stores.

7.3 Comparing Access

Restrictions to LC19 means that at the end of the project, LC18 data can be published as Linked Open Data, whilst LC19 data cannot. Although this has not affected the processes for creating RDF triples from the data, it does affect the publication and future possible re-use of the data from these two datasets in a fundamental way.

7.4 Comparing Live and Batch Processing

There are advantages and disadvantages to the live linked data production that complements the batch production of RDF triples (sections 4.3, 5.2, and 5.3; Table 2). *Ownership* (B-4, L+4) reflects the musicologists' desire for an authoritative and complete dataset [8]. Multiple dataset copies can cause uncertainty over version-control and availability. The reward structures of academia value publication of analysis and interpretation over dataset production, and can contribute to the delay in the open publication of the latter in order to ensure the former. *Access control* (B-5, L+5) refers to physical control over access, effectively describing the means by which the above considerations are enforced. The authoritative aspect of the *ownership* requirements could be met by ensuring that accepted academic usage of RDF always included provenance information. When included as part of attribution, datasets can be cited akin to the appropriate edition of a reference book.

LC18 and LC19 have authority files for people and venues. BMB provides a list of person names, with additional information. These were matched in a process combining automatic and human elements to create an internal linkage database [12]. Furthermore, the British Library has access to the *Concert Programmes Project* (CPP), which collates meta–information about archives including substantial authority files for people and venues [2]. These establish useful links within the CCP, and between person entities and VIAF[22], as well as geo–coding venues. Although full access was restricted, we could use CPP's persistent IDs, and extract sufficient information to match against *InConcert* authority files. The resulting linkages database has been used to augment human interconnection (Fig 10). Future availability of CPP as linked data will enable interconnection at a data instance level. The current process, employing simple lexical matching, could be extended with more sophisticated graph matching for entity identification.

[22]https://viaf.org/

Table 2: Comparison of batch (B) and live (L) RDF production

	Batch Workflows	Live Generation
storage	✓B+1: all the data together for processing	× L-1: distributed data hard to process
persistence	✓B+2: published snapshot, so external researchers able to say which version of data they are referring to (like published paper)	× L-2: data may change 'under your feet', could make external scholars' work harder. Data may even disappear when projects end.
liveness	✓B-3: not live when changes happen to original data sets	× L+3: as updates happen to scholarship, immediately reflected in public data
ownership	× B-4: loss of sense control for dataset owners	✓L+4: sense of control for dataset owners
access control	× B-5: potential loss of IPR, particularly if copied	✓L+5: access control possible

Figure 10: Links displayed with provenance

8. FUTURE WORK FOR LC18 AND LC19

Future work will focus both on project-specific tasks for LC18 and LC19, and on steps to support greater interlinking between the two datasets. LC18 will be improved by the minting of unique identifier for each Performance, as will LC19 for each new Record. Event-level linking in LC18 will increase the potential of connections between other datastreams holding historical data, and the anchoring of *InConcert* data within a context of historical events. The datasets already exhibit interlinking with one another, as well as linking out to external authority files. This information is visible via individual HTML views, but has not yet been implemented within the data views. The next stage of development will see further interlinking with online gazetteers.

All of the URIs used for the dataset will contain the prefix: http://datatodata.com/<uri>. The URIs schemas are illustrated in Table 3. As the LC19 data is currently secured, it is Linked Data but not Linked *Open* Data.

Both project datasets contain rich sections of information that are currently represented as plain-text strings. The scalable solution for providing robust semantic representations of this information is the development of a natural language processing tool, to pull out location, bibliographic, and prosopographical information as structured data. Such information facets will further improve opportunities for linking to external authorities such as MusicBrainz[23].

Table 3: schema for *InConcert* URIs

LC18
in-concert/lc18/concert/{id}
in-concert/lc18/place/{id}
in-concert/lc18/name/{id}
in-concert/lc18/type/{id}
in-concert/lc18/newspaper/{id}
in-concert/lc18/price_abv/{id}
in-concert/lc18/general_abv/{id}

LC19
in-concert/lc19/concert/{id}
in-concert/lc19/concert_item/{id}
in-concert/lc19/source/{id}
in-concert/lc19/source_item/{id}
in-concert/lc19/person/{id}
in-concert/lc19/venue/{id}
in-concert/lc19/work/{id}
in-concert/lc19/text/{id}

BMB
in-concert/bmb/entry/{id}
in-concert/bmb/page/{id}

9. RECOMMENDATIONS & CONCLUSION

In this paper, we have reported on different workflows to support the representation of tabular and relational data as Linked Data. We have compared the methods based on their level of user-control and on their dependency on the user's time, and verified the quality of the RDF triples generated. We have found that those workflows which require a greater investment of user-time initially produce a higher quality set of triples, and that even in the case of using an automated system such as D2R, some user-time is needed. Our recommendations are in line with the user's prior knowledge of existing systems, and level of technical competence. Scholars with limited technical expertise will benefit from the graphical user-interfaces of existing tools, but the process is likely to be time-consuming. Those familiar with the structure of the relational data may prefer the largely automated process of D2R. Finally, those with expertise in Semantic Web technologies may choose the JSON-LD approach to minimise the need for post-hoc tidying of the resultant RDF triples. All these approaches are most suitable for situations where the data is guaranteed to remain largely unaltered in the long term, as any changes to the raw data will necessitate the reiteration of the workflow in all these cases.

For both LC18 and LC19, human interpretation and evaluation informed the data interpretation and structuring. This first involves the musicologists creating the source data, requiring complex relational data modeling. The programmer's understanding of the data also plays a role, including enrichment of the data based on conversations with musicologists leading to revised versions of captured perspectives. A final level of interpretation occurs during ontological modeling, informed by the database structure and completed independently of the musicologist's input in our case.

[23]https://musicbrainz.org/

The use of common URIs is essential and supported by each of our workflows. Such URIs support large-scale processing (B+1 in Table 2), and authoritative reference (B+2), while maintaining a sense of ownership (L+4). Simultaneously, they enable up-to-date data publication (L+3) of *In-Concert* data as linked data, supporting unified access while enabling future reuse in external contexts.

10. ACKNOWLEDGEMENTS

This work was supported by the UK Arts and Humanities Research Council through the *Transforming Musicology* project (AH/L006820/1), part of the *Digital Transformations* theme, with additional support from the EPSRC *Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption (FAST IMPACt)* project (EP/L019981/1). The authors would like to thank their colleague David Lewis (Goldsmiths, University of London) for insights and advice regarding the SLICKMEM D2R workflow, and their colleagues Simon McVeigh (Goldsmiths), Rachel Cowgill (University of Huddersfield), Rupert Ridgewell (British Library), and Christina Bashford (University of Illinois at Urbana–Champaign) for their contributions, data, and analyses as the original creators of the *In Concert* project.

11. REFERENCES

[1] Konzertprogramm Austausch ('Concert Programme Exchange'), 1894–1944.

[2] Concert Programmes online database, 2016. accessed 3/1/2016.

[3] D. Bainbridge, X. Hu, and J. S. Downie. A Musical Progression with Greenstone: How Music Content Analysis and Linked Data is Helping Redefine the Boundaries to a Music Digital Library. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, DLfM '14, pages 1–8, New York, NY, USA, 2014. ACM.

[4] C. Basford, R. Cowgill, and S. McVeigh. The Concert Life in Nineteenth-Century London Database. In J. Dibble and B. Zon, editors, *Nineteenth-Century British Music Studies*, pages 1–12. Ashgate, Aldershot, 2000.

[5] M. Bay, J. A. Burgoyne, T. Crawford, D. De Roure, J. S. Downie, A. Ehmann, B. Fields, I. Fujinaga, K. Page, and J. B. Smith. Structural Analysis of Large Amounts of Music Information. *White paper, available via http://www. diggingintodata. org/Home/AwardRecipientsRound12009/StructuralAnalysisofLargeAmountsofMusic/tabid/179/Default. aspx*, 2009.

[6] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data–the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227, 2009.

[7] J. D. Brown and S. S. Stratton. *British musical biography: a dictionary of musical artists, authors, and composers born in Britain and its colonies*. SS Stratton, 1897.

[8] T. Crawford, B. Fields, D. Lewis, and K. Page. Explorations in Linked Data practice for early music corpora. In *Digital Libraries (JCDL), 2014*, pages 309–312. IEEE, 2014.

[9] D. De Roure. Executable Music Documents. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, pages 1–3. ACM, 2014.

[10] D. De Roure, G. Klyne, K. R. Page, J. P. Pybus, and D. M. Weigl. Music and Science: Parallels in Production. In *Proceedings of the 2nd International Workshop on Digital Libraries for Musicology*, pages 17–20. ACM, 2015.

[11] A. Dix, R. Cowgill, C. Bashford, S. McVeigh, and R. Ridgewell. Authority and Judgement in the Digital Archive. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, pages 1–8. ACM, 2014.

[12] A. Dix, R. Cowgill, C. Bashford, S. McVeigh, and R. Ridgewell. Spreadsheets as user interfaces. In *Proc. AVI2016*. ACM, 2016.

[13] A. Dix, A. Katifori, G. Lepouras, C. Vassilakis, and N. Shabir. Spreading activation over ontology-based resources: From personal context to web scale reasoning. *International Journal of Semantic Computing, Special Issue on Web Scale Reasoning: scalable, tolerant and dynamic*, 4(1):59–102, 2010.

[14] L. Dodds and I. Davis. *Linked Data Patterns*, chapter Identifier Patterns. 2011.

[15] L. Dreyfus and C. Rindfleisch. Using Digital Libraries in the Research of the Reception and Interpretation of Richard Wagner's Leitmotifs. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, pages 1–3. ACM, 2014.

[16] N. Harley and G. A. Wiggins. An Ontology for Abstract, Hierarchical Music Representation. 2015.

[17] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[18] S. McVeigh. Calendar of London Concerts 1750–1800. http://research.gold.ac.uk/10342/. Accessed: 2016-05-05.

[19] K. Page and P. Willcox. ElEPHãT: Early English Print in the HathiTrust, a Linked Semantic Worksets Prototype. 2015.

[20] S. Peroni and D. Shotton. FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17:33–43, 2012.

[21] Y. Raimond and S. Abdallah. The timeline ontology. *OWL-DL ontology*, 2006.

[22] Y. Raimond and S. Abdallah. The event ontology. Technical report, Citeseer, 2007.

[23] Y. Raimond, S. A. Abdallah, M. B. Sandler, and F. Giasson. The music ontology. In *ISMIR*, pages 417–422. Citeseer, 2007.

[24] D. Shotton and S. Peroni. FaBiO: FRBR Aligned Bibliographic Ontology, 2011.

[25] D. M. Weigl, D. Lewis, T. Crawford, and K. R. Page. Expert-guided semantic linking of music-library metadata for study and reuse. In *Proceedings of the 2nd International Workshop on Digital Libraries for Musicology*, pages 9–16. ACM, 2015.

[26] G. A. Wiggins, M. Harris, and A. Smaill. Representing Music for Analysis and Composition, 1989.